First Principles Selection of Social Media Visualizations

February 22, 2013

Sponsored by
Defense Advanced Research Projects Agency (DOD)
IPTO Office
(Dr. Rand Waltzman, Project Manager)

ARPA Order 4653-00

Issued by U.S. Army Contracting Command – Redstone
under
Contract no. **W31P4Q-12-C-0209**

Name of contractor:	InferLink Corporation
Principal Investigator:	Dr. Pedro Szekely
Business address:	2361 Rosecrans Ave. Ste. 348, El Segundo, CA 90245
Phone number:	(310) 341-2446
Effective date of contract:	June 22, 2012
Short title of work:	SocialViz
Contract expiration date:	February 28, 2012

1 Research goals

The goal of this STTR is to develop methods and tools for the dynamic, automatic generation of social network analysis (SNA) visualizations, based on established principles from cognitive science and cognitive neuroscience. Accomplishing this will allow analysts to better understand networks as well as make unique discoveries that are primarily facilitated by the visualization (i.e., "a picture is worth a 1000 words"). While tables of data are exacting and detailed, it nevertheless can be difficult, if not impossible, to spot a particular trend or anomaly until one actually sees it effectively illustrated. Node-link diagrams have been traditionally used in the context of social network analysis, but these are not always helpful, nor appropriate. If a visualization does not convey notable patterns and trends in a way easily perceived by humans, its key insights are effectively hidden. The problem is thus not the data, but the vehicle to illustrate that data. Our work on this project will focus on generating visualizations that adhere to cognitive first principles, so that key insights are not only communicated, but so that humans can also easily perceive them.

1 Financial Report

This section of the report will discuss the financial activity this period.

2.1 Incurred Expenses this Period

Direct Labor	$ 1,989.00
Overhead	$ 894.00
Subcontract	$14,550.00
Total Direct Costs	**$17,433.00**
G&A	$ 2,964.00
Total Costs	**$20,398.00**
Fee	$ 612.00
Total Cost + Fee	**$21,010.00**

2.2 Invoices this Period

Invoice 4005-07 was submitted in the amount of $12,499.00. The final invoice, 4005-08Z will be submitted upon submittal of the final report in the amount of $12,499.

2.3 Date Expenses will Equal Obligated Funding

The project was fully funded through February 22, 2012. InferLink did not require additional funding to complete the project.

2.4.1 Monthly Expenses

Jun 2012	Jul 2012	Aug 2012	Sep 2012	Oct 2012	Nov 2012	Dec 2012	Jan 2013	Feb 2013
$625	$13,458	$13,739	$13,482	$12,634	$11,067	$14,900	$11,631	$8,457

2.4.2 Total Expenses Planned for the Next 3 Quarters

There are no further expenses anticipated in connection with the phase one effort.

2.5 Projected Funding Increment

The project was fully funded through February 22, 2012. InferLink did not require additional funding to complete the project.

2.6 Issues or Concerns

No issues or concerns to report at this time.

3 Project progress

In this section, we discuss our progress this period and overall, throughout the entirety of the project. The two key goals we achieved this period were:

- Submission of Phase II proposal, which builds upon our work during this Phase I.
- Final report of the work we have done during Phase I (included below).

3.1 Phase I research

The goal of Phase I was two-fold:

1. Assess the feasibility of automatically generating insightful social network analysis (SNA) visualizations, based on established principles from cognitive science and cognitive neuroscience.

2. Design the architecture for implementing visualizations using these principles, and implement a proof of concept system that instantiates parts of these architecture to show feasibility of the approach.

Most SNA visualizations used today are "nodes and links" charts that encode information using the layout of the graph and the visual appearance of the nodes and links. For example, Figure 1 shows a chart of the neurological activity in the brain. This is a typical visualization that shows the density of connectivity between many entities. However, the heat map is ineffective because it is difficult to detect any pattern in the colors of the

nodes. While these types of visualizations are sometimes useful to convey patterns, often they are fail to inform because as the number of entities grows they tend to become messy "hairballs" where no patterns are apparent. In short, such visualizations fail to add insight or discovery value because they do not facilitate human cognition of patterns or anomalies.

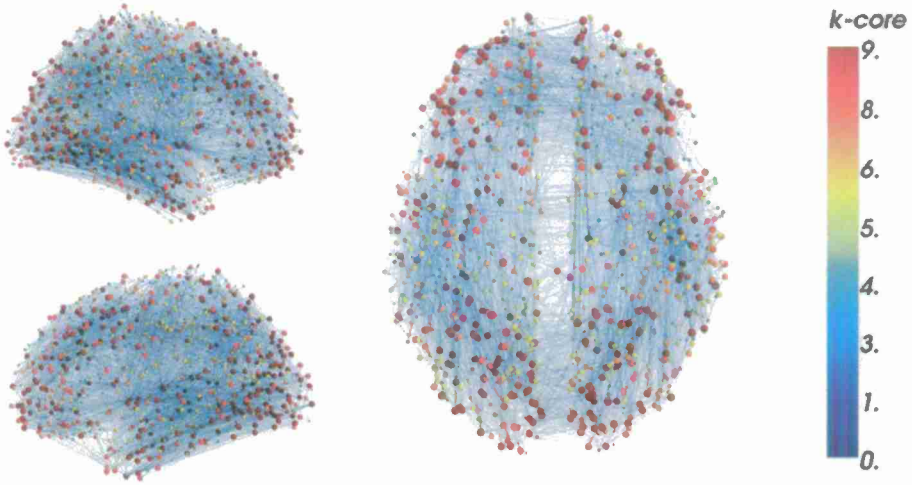

Figure 1: Hairball-like node-link diagram

Our objective in Phase I was to investigate visualizations for temporal social networks that do not use nodes and links. We focused on heat map visualizations that can aggregate information about a large number of entities without falling into the "hairball" trap.

3.1.1 Results from Phase I

The main result of our work is a general architecture for constructing principles-based visualizations of temporal social network datasets, as shown in Figure 2.

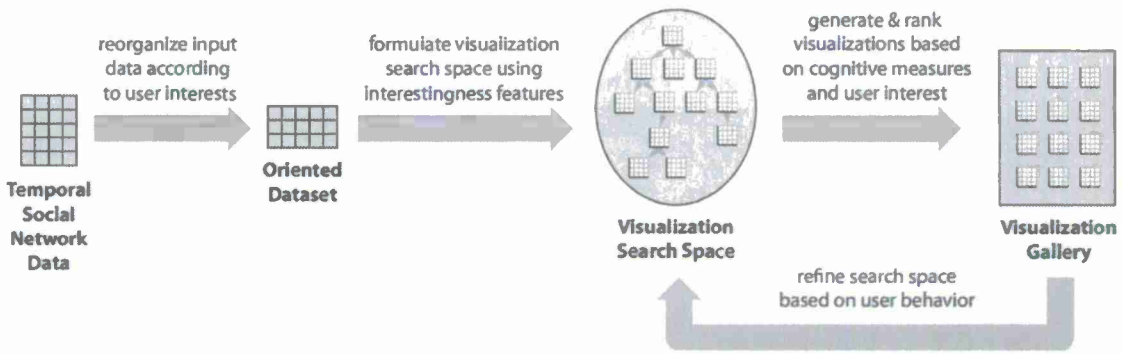

Figure 2: General architecture for principles-based visualization generation

The input to the system is a sequence of social interaction events, which we model as a sequence of tuples of the form (x, y, c, t), where x and y represent entities (the nodes), c is a vector of features about the interaction (the links) and t is a time-stamp. In Phase I, we

worked with a Twitter dataset (provided by Professor Macskassy), which consisted of tweets of people who reside in Middle Eastern countries. In this dataset, *x* and *y* represent people, and *c* is a vector containing a conversation identifier and other characteristics of the tweet such as it's length and the text of the tweet. The data includes roughly 2,500 nodes and 16,000 edges of network information, more nodes and links than can be sensibly visualized in a nodes and links visualization.

Professor Macskassy's goal was to understand how Twitter conversations relate to influence, diffusion, etc. Our goal was contribute new visualization technology to allow him to get insights into the data that he is not otherwise able to get from statistics alone. The corpus of data in the Twitter set he provided us helped us experiment with visualization approaches and techniques.

The key goal of our approach is to produce visualizations that show patterns that are not apparent in the data and that would not be revealed in typical node and link charts. To do this, we introduced the concept of an *orientation* that reorganizes the input data into a multi-dimensional dataset according to a dimension of interest to users. For example, in the Macskassy dataset we defined two orientations: a conversation orientation that organizes the data according to conversation identifiers so that we can visualize conversations rather than the individual tweets that form a conversation and that are scattered in time; a participant orientation that organizes the data according to participant so that we can visualize trends per participant.

In general, an orientation is a function of the raw (*x*, *y*, *c*, *t*) data where we first define a key, then aggregate the raw data according to this key and finally augment the data with derived variables. For example, in the conversation orientation we use the conversation identifier as the key (an element of *c*), and select all the tuples in the raw data with the same conversation identifier. These tuples will have different participants *x* and *y*, different time-stamps *t* and different characteristics *c*. In order to construct a multi-dimensional dataset that we can visualize we define new variables to aggregate the data. For example, the following are some of the variables defined in the aggregation step:

- **# participant**s: the number of participants in the conversation
- **duration**: duration of the conversation, form the earliest tweet to the latest tweet in the conversation.
- **start time**: the time-stamp of the earliest tweet in the conversation.

These new variables transform the original sequence of observations into a multi-dimensional dataset that can be visualized using many visualization techniques such as scatter plots and heat maps.

Finally we augment the resulting multi-dimensional dataset with additional derived variables that define features of the data that can lead to interesting patterns. For example, from the start time variable we define additional variables such as day of week, day of month, is weekend and hour of day. Using these additional variable we can construct

visualization that show patterns relating the length or number of participants in a conversation to the hour of the day of the week, the hour of the day, etc.

In Phase I we built orientations using custom software, but it became clear that it is possible to build a library of generally useful orientations that can be used for other types of temporal social network datasets. We plan to construct such a library in Phase 2 and we plan to build an interactive tool for interactively defining orientations in the option.

An orientation is a multi-dimensional dataset that defines a search space of visualizations. The search space is defined by the different subsets of variables that can be visualized and by the visual encodings that can be used to visualize the values of the selected variables. Figure 3 shows a gallery of visualizations for the conversation orientation of the Macskassy dataset constructed using our software. This gallery shows heat maps for a subset of the derived variables of conversation start times. Each heat map shows number of conversations with respect to selected variables. For example, the sixth visualization (*weekday* vs. *hourOfDay*) shows that our subject population engages in conversations more frequently in the evening and less frequently on weekends.

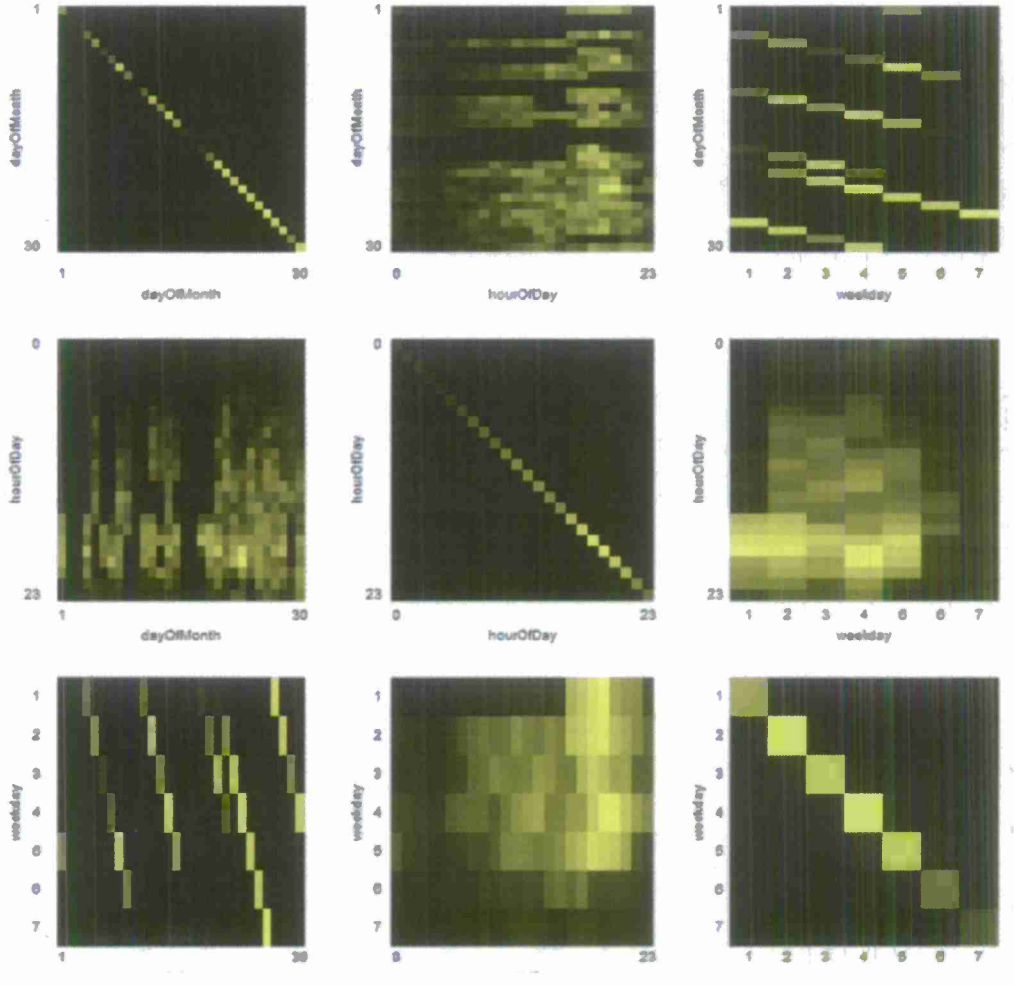

Figure 3: Heat map style visualization prototype developed during Phase I

However, the striking pattern in these visualizations is the unexpected "patches of black". For example, in the second visualization (*hourOfDay* vs. *dayOfMonth*) there are horizontal rivers of black showing that there were no conversations on some days of the month. The rivers of black extend to hours of the day when there were frequent conversations in most days of the month. Why? Perhaps Internet service was interrupted for extended periods in this part of the world. Further analysis revealed that the data collection software had crashed. The third visualization shows that during weekends the software had not been promptly restarted leading to data collection gaps for those days. So, in fact, the conclusion from the sixth visualization, namely that people tweet less during the weekends is probably incorrect.

The most interesting result of this part of the work is that it validated our hypothesis that visualizations can reveal patterns in the data that would be missed by social network analysis metrics. Of course, in some cases, it is easy to construct metrics that mathematically detect these types of patterns; the hard part is realizing that one should do so.

In Phase I, we built a prototype system that takes as input raw data (x, y, c, t), computes different orientations of the data, and automatically builds galleries of visualizations for pairs of variables in an orientation. Our software can build heat map visualizations such as the ones shown in Figure 3 for discrete variables, and scatter plots such as the one shown in Figure 4 (below) for continuous variables. Our visualization software generates visualizations using the *D3.js* (Data-Driven Documents) visualization software, an open-source, browser-based JavaScript library to generate SVG3 visualizations on the Web.

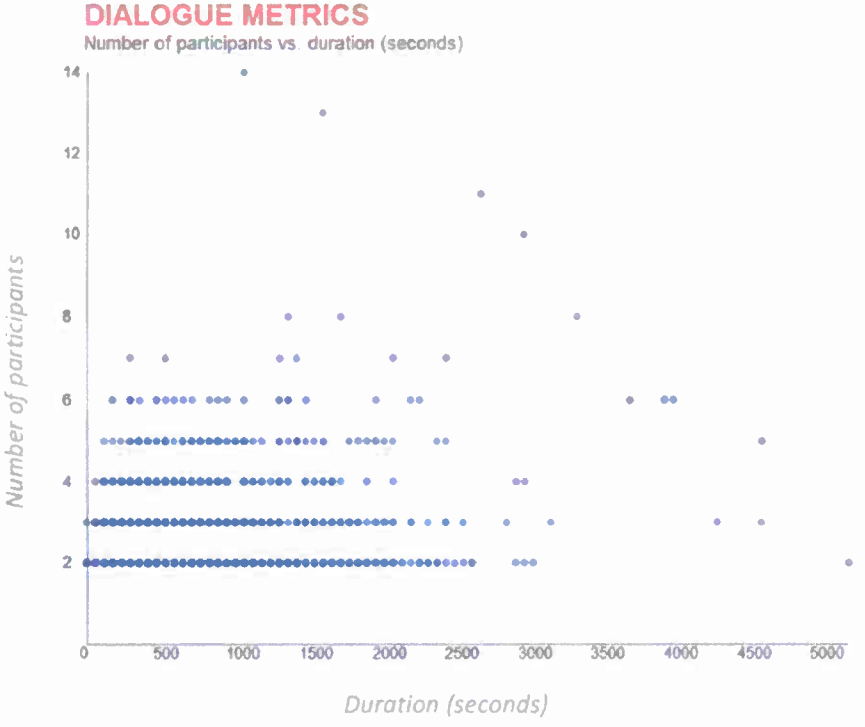

Figure 4: Scatter plot style visualization developer per our prototyping

The results from Phase I show that our approach is technically feasible and that it promises to lead to visualizations that suggest insights to analysts that they would not otherwise think about. Our prototype shows that our notion of orientations enables us to reorganize a social networks dataset into a multi-dimensional dataset that is amenable to a large variety of visualizations. It also shows that our approach can generate galleries of visualizations that show interesting patterns in the data.

One interesting aspect of data generation and manipulation that will have direct impact on the cognitive "accessibility" of the data is the clustering and ordering of the data. For example, in Figure 4, the pattern of dialogue duration range reducing as the number of participants grows is apparent from the figure. The organization of the data in way that reveals a trend is fortuitous in the above figure; there is a normal distribution over a continuous variable (time or number of participants). However, if the *y*-axis of the above graph had been about individual participants, there would be no natural order. While we could show the participants in any order, we potentially could facilitate better communication of trends by organizing the data in a sorted manner (longest duration to shortest, etc.). We might have also clustered the data into groups that reveal a trend (e.g., people who converse in mornings versus evenings, etc.). In short, grouping and ordering the data are key tools that help reveal patterns; however, we cannot use them alone, we must use them in ways that facilitate cognitive trend identification needs.

Even though our prototype does not yet implement cognitive measures to design visualizations, it shows a clear path forward to do so. The basic idea is to compute metrics on the resulting images to measure the interestingness of a visualization. For example, even a simple metric such as luminosity histogram would reveal that the first visualization in Figure 3 is interesting because the histogram has a sudden fall in the middle whereas the last visualization is uninteresting because the histogram is much more uniform.

Finally, one particular orientation we investigated in depth as part of our Phase I work, and in conjunction with the social dialogue data, relates to *participation*. More specifically, Macskassy was interested in looking at social connectivity patterns from the data set, illustrated across various other dimensions (such as time of day or day of the week).

We ran our system on the data set and identified the most popular graph formations, as shown in Figure 5 below. The figure shows the undirected graphs formations where n=3, 4, and 5 (the most popular node densities and ones of interest to Macskassy). In Figure 5, the solid lines indicate the most popular graph of that class (i.e., where class equals node size). The dotted lines highlight the next most popular graph type of that class. For example, the hub-spoke connected graph for n=4 is the most popular, with the fully connected graph being the second most popular.

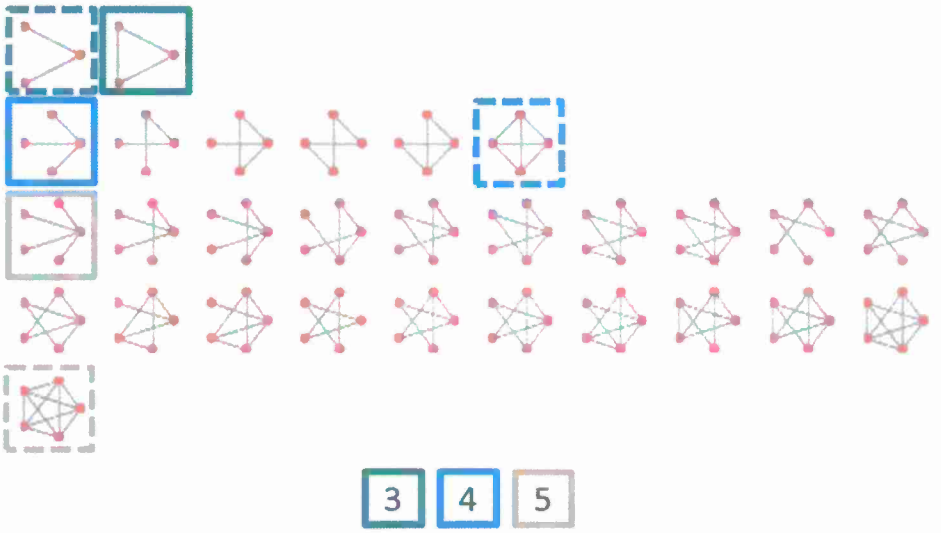

Figure 5: Connected graph formations for nodes of density 3, 4, and 5

We sent this data to Dr. Macskassy for further analysis and next steps. He has expressed interest in these findings. Our Phase II proposal supports the idea of defining orientations like this and exploring interesting visualizations once such data is paired with other contextual information (e.g., temporal data).

3.2 Deliverables this period

There were no deliverables this period.

3.3 Technology transition

There was no technology transition this period.

3.4 Publications this period

There were no publications this period.

3.5 Meetings and presentations this period

There were no meetings this period.

4 Project plans

The results of Phase I validate the vision for our approach and provide a path for achieving it. Specifically, our key objectives of our Phase II proposal are as follows:

- **Automatically generate insightful visualizations based on perceptual and cognitive principles.** Our objective is to create visualizations that go beyond the traditional node and link visualizations, and to create visualizations that enable researchers and analysts to see patterns in the data that they didn't know exist. To this end, our system will transform the temporal social network data in different ways, which we call orientations, detect interesting patterns in the data, and automatically generate visualizations that convey these patterns clearly. To do so,

our system will generate galleries of visualizations, each one highlighting an interesting pattern in the data, and each one designed according to cognitive and perceptual principles so that users can easily see and understand the patterns.

- **Model human interest in visualizations and combine with cognitive principles.** One of the key things we learned in our Phase I work was that the space of visualizations, even if filtered by those that look interesting may still be enormous. The search space explodes as more social dialogue attributes are added to the data set - there are simply too many ways that attributes could be combined, many of which may lead to potentially interesting insights. One way to refine search further is by understanding evolving human interests and continually refining our model to suit that interest. This is a common technique that is successfully used to refine search in large online retail systems, for example. Thus, our second objective in Phase II is to model human interests, such as those found in recommender systems and use this knowledge to further refine the user's search through the space of visualizations. By combining cognitive models and user models, we can build a hybrid evaluation system that targets both needs, just as hybrid recommenders have been successfully found to target both content-based and collaborative filtering algorithms.

- **Deployment and evaluation of fully integrated system.** A third objective in a Phase II is to combine our cognitive measures and human interest modeling technologies into a single, open-source system that can be used by anyone who wishes to understand temporal social network phenomena. Meeting this objective will require that we build the remaining components to deliver an open source, Web-based visualization portal, which are inspired by our prototyping in Phase I. Just as we worked with Dr. Macskassy in Phase I, we also intend to work with social network computer scientists like Dr. Kristina Lerman and others working in conjunction with the Social Media in Strategic Communication (SMISC) effort, who have significant experience producing social network analyses, as well as operational experts. These latter subject matter experts (SMEs) will provide insight into how our tool might affect operational analysis and will help guide our iterative design. Both types of experts will help us validate our technology.

At this point, the Phase I project is complete and we have no further technical development plans. We have submitted a Phase II proposal and are waiting on its review.